DOCUMENT RESUME

ED 301 573 TM 012 375

AUTHOR Mouw, John T.; Perkins, Kyle
TITLE A Framework for Conducting ESL/EFL Construct Validation Studies.
PUB DATE 86
NOTE 20p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Construct Validity; Elementary Secondary Education; *English (Second Language); *Generalizability Theory; *Latent Trait Theory; Learning Readiness; *Research Methodology; Test Reliability; Test Use; *Test Validity
IDENTIFIERS Pearson Product Moment Correlation; Rasch Model; *Validity Research

ABSTRACT

The purpose for which a test is used and the examinees' stage of learning are two anchor points that are incorporated into a suggested framework for conducting construct validation studies for tests of students with English as a second language (ESL) or English as a foreign language (EFL). The framework includes the use of generalizability theory, construct definition, and factor analysis. Two procedures are discussed in terms of the means by which they confound reliability and validity. Difficulties with correlational evidence to support construct valid test score interpretations are also discussed. Problems with product-moment correlations, the Rasch model, and a construct validation technique proposed by G. H. Henning are singled out for particular criticism. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301573

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KYLE PERKINS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

John T. Mouw and Kyle Perkins

Southern Illinois University

TM012 375

BEST COPY AVAILABLE

A Framework for Conducting ESL/EFL Construct Validation Studies

Abstract

The purpose for which a test is used and the examinees' stage of learning are two anchor points which are incorporated into a suggested framework for conducting ESL/EFL construct validation studies. The framework includes the utilization of generalizability theory, construct definition, and factor analysis. Two procedures are discussed in terms of how they confound reliability and validity. Difficulties with correlational evidence to support construct valid test score interpretations are also discussed.

Introduction

Our purpose in this paper is to introduce to the ESL/EFL community a structure or framework for conducting construct validation studies and to motivate that framework. Our suggested framework includes the utilization of (1) generalizability theory, (2) construct definition, and (3) factor-analytic techniques. We will also discuss two procedures which seem to us to confound reliability and validity. Although the confounding of reliability and validity is not our thesis, we will address the issue, because it is more than tangentially related to our principal topic--construct validity.

Definitions of construct validity

We begin our presentation with some definitions of construct validity. The American Psychological Association Standards (American Psychological Association, 1974) defined a psychological construct as "a theoretical idea developed to explain and to organize some aspects of existing knowledge" (p.29). According to Cronbach and Meehl (1955), the process of construct validation occurs

> ....when an investigator believes that his (or her) instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim (p. 255).

Our working definition of construct validity then is the extent to which a test measures the trait it was designed to measure. There seems to be a variety of terminology used for the word trait. In our review of the literature we have encountered the following terms in addition to trait: knowledge, competence, ability, construct, and proficiency. John B. Carroll (personal communication) has suggested that the terms

ability and proficiency be used in this context rather than trait. To Carroll, ability is a loaded term, because it connotes a more permanent, fixed capacity; proficiency, the preferred nomenclature, can be described in terms of what kinds of tasks the examinee is able to perform, and one can describe the tasks in terms of outcomes, requirements, and stimulus conditions. Carroll ruled out the use of trait, because if one measured ESL/EFL proficiency after three months of training, one would be measuring a state of learning at a particular time (not a trait) which, hopefully, would change after yet another three months of training.

Test developers validate the desired interpretations from test scores by a chronological progression of activities, according to Hambleton and Swaminathan (1985):

> A theory that incorporates the trait measured by the test is formulated. Hypotheses derived from the theory about how the ability scores should function (i.e., what should the ability scores correlate with? What factors affect ability scores?) are stated. These hypotheses may relate to evidence that comes in the form of (1) analyses of test content, (2) reliability studies, (3) correlations of ability scores with other measures, (4) experimental studies, (5) prediction studies, (6) multi-trait multi-method investigations. The necessary data for the investigation of the hypotheses are collected and the results analyzed. The results are interpreted in relation to the hypotheses (pp. 71-72).

An anchor point for our proposed structure for conducting ESL/EFL construct validation studies is Nitko's (1983) position that validity is not a characteristic of tests. Whether a test is valid depends on

the purposes for which it is used and in what contexts. A logical inference is that a particular test may be valid for a specific purpose but not valid for other purposes. In a volume devoted to the measurement of reading, Farr and Carey (1986) stated that "the most important validity question is whether a particular test meets a user's specific information needs; test validity relates to how valid a test may be for a given purpose" (p. 147). To illustrate the point, Farr and Carey (paraphrasing from Nitko) offered several examples of validity questions that depend on the testing purposes:

> Content validity: To what extent are the items a representative sample of content (behavior) from the test publisher's domain? To what extent are the psychological processes required by the test items representative of the processes required by the test developer's domain?
>
> Curricular relevance: To what extent are the test items and their required psychological processes representative of the domain as we define it (e.g., our school curriculum, our view of intelligence, and/or our view of a certain personality trait?)
>
> Construct validity: To what extent can the test scores be interpreted as meaningful beyond the mere level of items to the level of construct interpretations which are believed to underlie observed performance on the test items?
>
> Criterion oriented validity: To what extent can scores on the test be used to infer a person's standing on a second measure (criterion)? a. To what extent can the scores be used to select persons for jobs, schools, etc.? What is the magnitude of the

> error? b. To what extent can the test be used to assign pupils to different levels of instructions (p. 147).

We believe that these positions taken by Nitko and Farr and Carey clearly render invalid any claim that construct validity is omnibus, subsuming all other types of validity. Also inherent in this discussion are the question "Does a test actually tap a particular ability or behavior?" and the premise that one test reflects only one behavior at the time of testing.

In addition to the question of a test's being valid for a given purpose, we must also consider the fact that the examinees' stage of learning must be included in our design of ESL/EFL validation studies. Messick (1984) in a discussion of how achievement measures should be cast in developmental terms noted

> The constructs emphasized differ depending on whether the students assessed are at beginning, intermediate, or advanced levels of learning of the subject area. Because in beginning or low-level achievement a major issue is the acquisition of a critical mass of information on the subject, the constructs might include information retrieval and interpretation relative to the student's naive knowledge structure and beliefs about the subject area; the education of substantive relationships to augment, transform, and refine the naive structure; and, specific skill applications. At intermediate levels of achievement, the constructs might stress not only the structuring of newly acquired knowledge but also the structuring of knowledge-dependent skills and more general cognitive abilities for substantive understanding and directed thinking in the area. In advanced or expert-level achievement, the emphasis

might be on representational and restructuring skills in concert with the flexible application of multiple specialized schemas for problem representation and solution (p. 216).

Generalizability theory

We have presented arguments that any construct validation study must include, among other considerations, the purpose for which the test was intended and the examinees' stage of learning at the time when the test was administered. We believe that generalizability theory can accommodate all the necessary and sufficient conditions for construct validation.

Generalizability theory (Brennan, 1983) recognizes and accommodates multiple sources of variation in assessment; it utilizes analysis of variance to estimate variance components for different effects to evaluate measurement dependability. An observation on a measurement object is presumed to be sampled from a universe of observations. The set of all conditions under which an observation is made is called a facet. The facets which are investigated define a universe of admissible observations. From the universe of generalization comes the expected value of the observed score which is called the universe score, and universe scores are estimated by the mean from a sample of observations. The observed score is treated as an estimate of the universe score, and a generalizability coefficient describes the dependability of generalizing from observed scores to universe scores.

What we find particularly appealing about using generalizability theory for construct validation studies is that for each facet there exists a universe of conditions (finite or infinite) from which the conditions for a particular study may be drawn. A broad spectrum universe of conditions allows one to take cognizance of the test's

purpose and the examinees' stage of learning.

Kane (1982) noted that the empirical testing of invariance laws is necessary in establishing the validity of a measurement procedure. He defined validity "in terms of invariance over a universe of generalization and (validity) is associated with the 'meaning' of an attribute in the narrow sense of denotation" (p. 153).

Kane used generalizability theory to develop a multifacet sampling model to measure dispositional attributes:

> A dispositional attribute is defined in terms of a universe of generalization. The "true" value of the attribute is the universe score, equal to the expected value over the universe of generalization. Measurement involves inferences from observations to the expected value over the universe defining an attribute, and a measurement procedure is valid to the extent that its inferences about universe scores are accurate (p. 154).

Kane provided the following validity index which reflects the expected agreement between observed scores and the value oof the attribute:

$$\rho^2 (X_{oI*R}, \mu_o) = \frac{s^2 (o)}{s^2 (o) + s^2 (oI) + s^2 (R)}$$

Kane also provided a formula for a validity coefficient corrected for attenuation:

$$\rho^2 (\mu_o *, \mu_o) = \frac{\rho^2 (X_{oI*R}, \mu_o)}{\rho^2 (X_{oI*R}, \mu_{o*})}$$

Where

ρ^2 = an index of the squared correlation between observed scores and universe scores.

X_{oI^*R} = an observed score for a measurement procedure with the i facet standardized to I*.

μ_o^* = the expected value of the observed score over repeated application of the standardized measurement procedure.

o = the object of measurement.

I = a sample of n_i conditions of the i facet.

R = a sample of n_r replications for each condition of the i facet.

μ_o = the universe score for the object of measurement.

Construct definition

Embretson (1983) recommended a paradigm shift in construct validation from functionalism to structuralism--a shift "from explaining antecedent/consequent relationships to explaining performance from the systems and subsystems of underlying processes" (p. 179). She proposed that two separate concepts be addressed by construct validation research: construct representation and nomothetic span:

> Construct representation is concerned with identifying the theoretical mechanisms that underlie task performance. In short, the goal of construct representation is task decomposition. In the information-processing paradigm, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge stores that are involved in performance.
>
> Nomothetic span refers to the network of relationships of a test to other measures. Nomothetic span is supported jointly by the strength, frequency, and pattern of significant relationships with other measures, such as other traits, criterion measures and so forth. It is assessed by individual

differences data (p. 180).

Embretson used multicomponent latent trait models to analyze the concepts of construct representation and nomothetic span. Multicomponent latent trait models link task performance to traits and characteristics of the task to item difficulty and person ability. Embretson also suggested using structural equation analysis to examine the nomothetic span of the strategies and the metacomponents of strategy application for the purposes of estimating the contribution of the constructs to account for individual differences, to evaluate the contributions of the constructs to test validity, to estimate prediction weights, and to provide a significance test of each prediction weight and a goodness-of-fit test for the entire model.

While Embretson's proposal is indeed impressive, one's enthusiasm for its entire application to ESL/EFL language proficiency testing must be leavened with caution for at least three reasons. First, it may not possible to design as many tasks and measures for ESL/EFL language tests as Embretson's model requires. An earlier Embretson (Whitely) (1980) paper on aptitude test validity examined some six components, two content factors, and three process factors. Second, John B. Carroll (personal communication) has suggested that language proficiency test items might not be multicomponential, and if they are, the magnitude of labor necessary to demonstrate multicomponentiality would not be warranted. Third, John A. Upshur (personal communication) doubts the existence of any language test tasks for which the probability of correct performance would depend upon the joint probability of performing some set of component tasks.

However, it seems to us that there is one aspect suggested by the Embretson model which has potential utility in ESL/EFL construct validation studies--construct definition. Following the paradigm set out by Stenner, Smith, and Burdick (1983), Perkins and Linnville (in preparation) are conducting a validation study of a Michigan Test of English Language Proficiency vocabulary subtest. The difficulty of the Michigan vocabulary items is attributed to the following characteristics of the stimulus words: (1) number of letters, (2) number of syllables, (3) log frequency-the common logarithm of the frequency of a word's appearance in large samples of written material, (4) dispersion-the degree to which the word is likely to be encountered across different content areas, (5) whether the word is concrete or abstract, (6) a word's emotive value, and (7) the number of attributes possessed by an object that match attributes in the word's definition. The values for each of the characteristics will be submitted to intercorrelations and raw score regression equations to determine how much of the observed score variance is accounted for in the item scale values by the construct specification equation.

Factor analytic studies of writing ability

There has been debate in ESL/EFL language testing about whether language proficiency is unidimensional (Oller, 1979; Oller and Perkins, 1978, 1980) or multidimensional (Farhady, 1980; Oller, 1983). The same field of inquiry could be transferred to ESL/EFL composition evaluation in order to determine whether empirical essay data show evidence for a single factor or a multifactor model of judged writing ability. An ancillary issue is whether this factor (these factors) relate to objective test measures of writing ability (cf. Rock, 1986). The principal questions are these: how many constructs underlie writing ability and to what extent do objective tests of writing account for variance in these constructs?

These questions could be answered by applying confirmatory factor analytic techniques to holistically scored essay data and multiple-choice objective test measures. Some potential multiple-choice tests that might have usefulness for such an enterprise include Test of Ability to Subordinate (TAS), Test of Standard Written English (TSWE), and Revision and Editing Test (RET), although the latter two tests would have to be adapted for ESL/EFL populations because they are first language tests.

The TAS is a 50-item sentence-combining test which is said to measure specific grammatical points and is recommended for diagnosis and placement purposes. Kunz (1980) reported that the TAS and cloze scores correlate significantly with the Michigan test and ESL placement scores.

The TSWE is a thirty-minute 50-item multiple-choice test which evaluates an examinee's ability to recognize standard written English. The purpose of TSWE scores is to place students into appropriate freshman composition courses. It has been administered as a part of the Scholastic Aptitude Test. Two types of questions are included--those on usage and those on sentence correction.

The RET is an exit exam designed to assess an examinee's ability to revise and to edit; the test was described in Matalene (1982).

To conduct such a factor analytic study on the number of constructs underlying writing ability, one could use LISREL (Joreskog and Sorbom, 1985) to test a series of factor models (i.e., the maximum likelihood estimation of the factor models). The estimation process begins with a single factor model of writing proficiency and then proceeds to more complex models, e.g., two, three, or four factors. Goodness-of-fit measures would determine how well the data fit the model. Of particular

interest would be the pattern of factor loadings of the objective measures. Studies of this type could shed light on whether there is only a dominant, general ESL/EFL writing proficiency factor (construct), a dominant factor with definable subfactors, or multiple definable factors.

Reliability and validity confounding

In our review of the literature we noted at least two instances where reliability and validity have been confounded. Perhaps a reason for this confounding is that both reliability and validity concern error, but different kinds of error. Kane (1982, p. 126) noted that reliability is associated with random errors which vary from one observation to another, while validity is associated with systematic errors which are constant for a series of observations.

A procedure which seems to us to confuse reliability and validity is a construct validation technique proposed in Henning (in press) that relies on item statistics. Henning has suggested that point-biserial correlations be calculated between each item and the total test scores. Another set of point-biserial correlations between each item from the test under consideration and a criterion test (different construct) are examined. The first set of correlations is corrected for part-whole overlap. An item is said to demonstrate construct validity if it is correlated more highly with its own test total than with the external criterion measure.

At first blush Henning's propsal is intuitively appealing, but there are some vexing problems. First, the intercorrelations of items with the total test (corrected for overlap) is at the very foundation of the KR-20 and Cronbach's alpha, both of which are estimates of reliability. Second, the comparison of such indices with those for an outside criterion adds only an indirect estimate of validity. Third,

the external criterion measure might be relevant or irrelevant. Fourth, John B. Carroll (personal communication) has pointed out that if an item is correlated more highly with its own test total than with the external measure, this might show only that the test measures something of completely trivial importance. Fifth, point-biserial correlations have well-known limitations: (1) it is assumed that there is one variable which is scored dichotomously, yet the probability of an examinee's answering correctly is continuous and (2) it is also assumed that there exists continuous, interval data, which may not always be the case.

There are also problems with product-moment correlations. Positive results can be used as evidence to support construct valid test score interpretations; however, negative evidence can be interpreted in a variety of ways: (1) the measures were unreliable; (2) the theory is faulty; (3) the test does not measure the trait of interest; (4) the hypotheses were incorrect; (5) the method of correlation was incorrect--for instance, the regression might be non-linear forcing the choice of another, more appropriate correlational technique (Thomas M. Haladyna, personal communication).

We also believe that the Rasch model is inadequate for construct validation studies because it confuses reliability and validity. Wright and Stone (1979) suggest that it is possible to use Rasch one-parameter item response test statistics "to show how a calibrated item can be used to define a variable and ... to find out whether the resulting operational definition of the variable makes sense" (p. 83). The first step is to determine, from the item difficulties, the degree to which the data have given a direction to the variable. The items are arranged according to their item difficulties at their calibrated positions along the line of the variable. After the items have been so arranged, it is possible to

ascertain which items define the easy end of the variable, which items define the difficult end, and where the wide undefined gaps are, i.e., areas of difficulty for which there are no items. Such a procedure rests on the assumption that the data on which the item is based are unidimensional and assess only one construct.

We believe that a good fit between the test data and the chosen model is a matter of reliability, a sort of internal consistency between the model and the data set and has, consequently, nothing to do with what the test measures, i.e., construct validity, what we are on about in this paper. If a data set is truly unidimensional and there is goodness-of-fit between the data and the item response model, the items can be said to measure a common trait or construct.

Hambleton and Swaminathan (1985) offered more telling commentary:

> Wood (1978) highlights the problem of model-data fit and what a test measures. He demonstrated that coin-flipping data could be fitted to an item response model. In the example provided by Wood (1978), a candidate's score was the number of heads coming up in ten tosses of a coin. For this example, Wood (1978) was able to obtain both item parameter and ability estimates. The fit between the data and the one-parameter model was very good even though the underlying trait was both invalid and unreliably measured! Therefore, when ability scores are obtained, and before they are used, some evidence that the scores serve their intended purpose is needed (p. 71).

Implicit in any consideration of construct validation, the procedure employed notwithstanding, must be a concern with construct (ability) variance. Before one can begin to deal with construct variance, one must identify and isolate test variance which is caused by the underlying

constructs or abilities, by method effects, by instruction or training, by exposure to the language, and by measurement error (cf. Upshur and Homburg, 1983; Purcell, 1983).

Conclusion

The issue of construct validity is important as evidenced by the amount of attention given to it in the measurement texts and the literature. Secondly, the study of construct validity is difficult judged by the variety of techniques developed to obtain a clear picture of that elusive thing called a construct. In this paper we have suggested a framework for conducting ESL/EFL construct validation studies, and we have attempted to incorporate the purpose for which the test was designed and the stage of the examinees' learning into that framework.

References

American Psychological Association. 1974. Standards for educational and psychological tests. Washington, D.C.: Author.

Brennan, R. L. 1983. Elements of generalizability theory. Iowa City, Iowa: The American College Testing Program.

Cronbach, L. J., and Meehl, P. E. 1955. Construct validity in psychological tests. Psychological Bulletin 52, 281-302.

Embretson, S. E. 1983. Construct validity: Construct representation versus nomothetic span. Psychological Bulletin 93, 179-197.

Farhady, H. 1980. Justification, development, and validation of functional language testing. Ph.D. dissertation, University of California, Los Angeles.

Farr, R., and Carey, R. F. 1986. Reading, What can be measured? Second edition. Newark, Delaware: International Reading Association, Inc.

Hambleton, R. K., and Swaminathan, H. 1985. Item response theory. Principles and applications. Boston, Massachusetts: Kluwer.Nijhoff.

Henning, G. H. In press. Language test development. Rowley, Massachusetts: Newbury House Publishers, Inc.

Joreskog, K., and Sorbom, D. 1985. Lisrel 6, Analysis of linear structural relationships by the method of maximum likelihood. User's guide. Mooresville, Indiana: Scientific Software, Inc.

Kane, M. T. 1982. A sampling model for validity. Applied Psychological Measurement 6, 125-160.

Kunz, L. 1980. English as a second language. In A. Trillin and Associates (Eds.), Teaching basic skills in college (pp. 145-207). San Francisco, California: Jossey-Bass.

Matalene, C. 1982. Objective testing: politics, problems, possibilities. College English 44, 368-381.

Messick, S. 1984. The psychology of educational measurement. Journal of Educational Measurement 21, 215-237.

Nitko, A. J. 1983. Educational tests and measurement: An introduction. New York: Harcourt Brace Jovanovich, Inc.

Oller, J.W., Jr. 1979. Language tests at school. London: Longman Group Limited.

Oller, J. W., Jr. (Ed.). 1983. Issues in language testing research. Rowley, Massachusetts: Newbury House Publishers, Inc.

Oller, J.W., Jr., and Perkins, K. (Eds.). 1978. Language in education: Testing the tests. Rowley, Massachusetts: Newbury House Publishers, Inc.

Oller, J.W., Jr., and Perkins, K. (Eds.). 1980. Research in language testing. Rowley, Massachusetts: Newbury House Publishers, Inc.

Perkins, K., and Linnville, S. E. In preparation. A construct validation study of a Michigan Test of English Language Proficiency vocabulary test.

Purcell, E. T. 1983. Models of pronunciation accuracy. In J. W. Oller, Jr. (Ed.), Issues in language testing research (pp. 133-153). Rowley, Massachusetts: Newbury House Publishers, Inc.

Rock, D. 1986. Essay tests: What is being measured, how does it differ from what is being measured by multiple-choice tests? Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April.

Stenner, A. J., Smith, M., and Burdick, D. S. 1983. Toward a theory of construct definition. Journal of Educational Measurement 20, 305-316.

Upshur, J. A., and Homburg, T. J. 1983. Some relations among language tests at successive ability levels. In J. W. Oller, Jr. (Ed.), Issues in language testing research (pp. 188-202). Rowley, Massa-

chusetts: Newbury House Publishers, Inc.

Whitely, S. E. 1980. Modeling aptitude test validity from cognitive components. Journal of Educational Psychology 6, 750-769.

Wood, R. 1978. Fitting the Rasch model-A heady tale. British Journal of Mathematical and Statistical Psychology 31, 27-32.

Wright, B. D., and Stone, M. H. 1979. Best test design. Rasch measurement. Chicago, Illinois: MESA Press.